Time-Series Cross-Sectional Analysis

Jan Rovny

4/2/2020

This tutorial is based on what I learned from lectures and discussions with James A. Stimson at UNC Chapel Hill. Most of what I know about time-series cross-section is thanks to him. This tutorial borrows some R code from the work of Oscar Torres-Reyna

Introduction

As social scientists, we are often interested in studying phenomena that occur over time, but in different units of observation, such as countries. It effectively means that we have a number of parallel time-series in different sections to analyze. This type of analysis is called *time-series cross-section* (TSCS). Time-series cross-sectional analyses thus contain time-ordered observations across different units.

When we carry out TSCS, we assume that the same type of causal process occurs in all our units. This is to say that we have the same basic theoretical expectations about the relationships between variables in all our units. A formal way of saying this is that we expect theoretical *unit homogeneity.* If this is not the case, and we believe that different mechanisms are at play across the units, we should probably not combine our units into a common analysis, or we should try to identify contextual variables that can account for this variation. This reflects an inheret tension in social science. On the one hand, we wishe to be case sensitive, and account for important variation across units (countries, regions etc.). On the other hand, we wish to develop more general theories that apply across cases, not just in one unit.

Data structure in TSCS

When we combine time-series across units, we get a particular data structure that looks like this:

Unit	Time
1	1
1	2
1	3
1	...
1	T
2	1
2	2
2	3
2	...
2	T
...	...
N	1
N	2
N	3

Unit	Time
N	T

This means that number of units is denoted with N, and time-periods are denoted with T. Our total number of observations is is thus $N * T$.

If we consider a basic regression model:

$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{u}$

Then in TSCS:

$\boldsymbol{y}$ is a vector of length $N * T$

$\boldsymbol{x}$ is a matrix with $N * T$ rows and $k + 1$ columns, where k is the number of predictors in the model, and the additional 1 is for the constant.

$\boldsymbol{\beta}$ is a vector of length $k + 1$, where k is the number of predictors in the model, and the additional 1 is for the constant.

$\boldsymbol{u}$ is a vector of disturbances of length $N * T$

Problems in TSCS data

Given the nature of time-series cross-sectional data, we face a number of problems that we need to consider when we carry out our analyses.

1) Each unit is likely to have some unit-specific context that is not modelled. This should speak directly to country expects, keenly aware of the specificities of a given place, its history, customs, dynamics. More generally, each unit has some set of idiosyncracies that are not considered in the model, and thus are projected into a unit-specific constant. As a result, we should add a constant C_i to $\boldsymbol{X\beta}$. If this idiosyncracy is omitted, it amounts to omitted variable bias, which can be a big problem, biasing estimates and standard errors.

2) Since we are dealing with time-series, TSCS is likely to encounter autocorrelation. The error-aggregation processes that we studied in Time-Series 1 are likely to be present, biasing our estimates of standard errors, and undermining our error-testing.

3) Given the differences in units discussed in 1), we are likely to also see unequal variances between them, meaning that $\sigma_i^2 \neq \sigma_j^2$ for units i and j. For example, government spending in the United States and Eritrea is likely to have different variances. This heteroscedasticity, as we know, would also bias our standard errors.

4) Errors may be correlated (not-independent) across spatially proximate units because they are related to each other, biasing standard error estimation.

5) Finally, common exogenous events occurring at the same time (think Coronavirus crisis across most countries of the world in 2020) could make units at the same time not independent of one anther, lead to correlation of errors across units at the same time, again biasing our standard errors.

All these problems seem, quite frankly, overwhelming. But let's put things into perspective, and consider which problem is most threatening to our ability to estimate a valid model. Clearly, it is problem 1). Why? Because only problem 1) may leading to bias in the estimates – that is, in our ability to assess the directions and strengths of the effects of our predictors. The other problems only threaten our standard errors, and thus our ability to test whether the estimates are statistically significant.

Unit effects

Problem 1) above engages a major issue in TSCS analysis, known as *unit effects*. Although we assume that there are homogeneous causal processes at play across our units, we are aware of the fact that historical and contextual idiosyncracies of our units are likely to lead to different levels of the dependent variable y across our units. For example, when studying economic development of European Union countries between 2010 and 2020, we know that these countries have widely diverging past experiences of economic development, different institutions etc., in short, their levels of economic development differ significantly – these are unit effects. Unit effects suggest that 'all else is not equal' or *ceteris non paribus…*

How can we assess whether our data have significant unit effects? We can at first simply look at mean levels of the dependent variable across the units. A more formal test of unit effects is an analysis of variance (ANOVA) by units.

Let's consider an example of eleven eastern European countries over the the course of their post-communist history. We are interested in explaining unemployment levels across the region. The dependent variable in the data is `unempl`, the unit variable is country, coded into a simplified country id variable `cid`. The time variable is `year`.

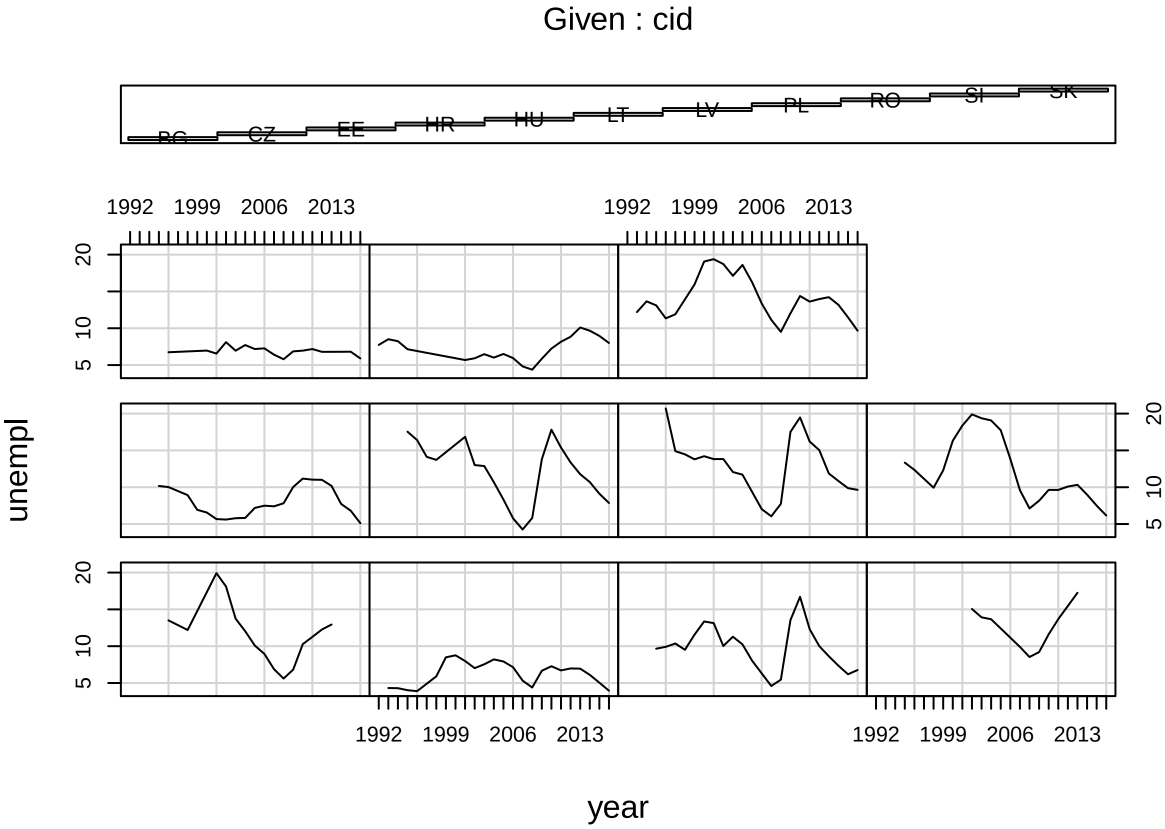
Let's first load the data and all the R libraries we will need

```
#Import data
library(rio)
D<-import("https://jan-rovny.squarespace.com/s/WB.csv")

#Load all necessary libraries
library(tidyverse) # Modern data science library
library(plm)       # Panel data analysis library
library(car)       # Companion to applied regression
library(gplots)    # Various programing tools for plotting data
library(tseries)   # For timeseries analysis
library(lmtest)    # For hetoroskedasticity analysis
```

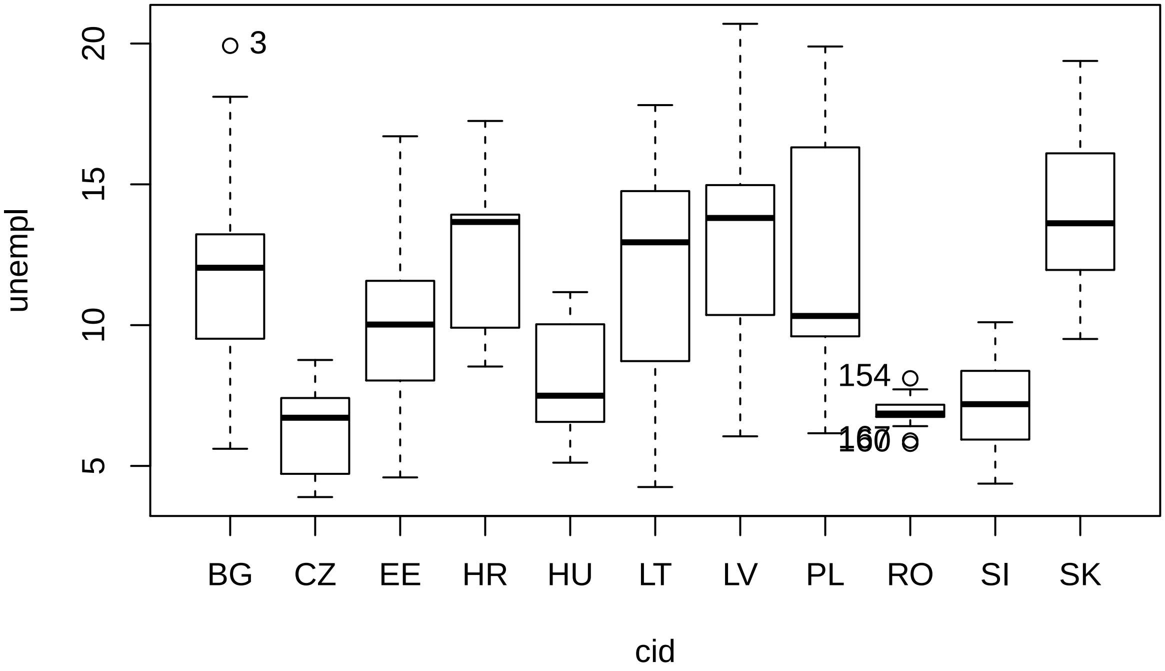
Now, let's declare the dataset TSCS data, and first view the series of the dependent variable across the units:

```
D<-pdata.frame(D, index=c("cid", "year")) #declare TSCS data, identify unit and time variables
coplot(unempl ~ year | cid, type="l", data=D) #plot unemployment by year and cid
```



The figure above nicely demonstrates the fluctuations of the dependent variable over time and across our eleven countries. Looking carefully at the figures suggests that different counties are likely to have different overall (mean) levels of unemployment. Let's take a more careful look at the distributions of the dependent variable across the units:

```
scatterplot(unempl ~ cid, data=D) #view distribution of unemployment by cid
```



```
## [1] "3"   "160" "167" "154"
```

The above figure shows two things. First, and most important, levels of unemployment clearly differ starkly accross the countries. The Czech Republic and Slovenia have low levels, while Latvia and Slovakia have much higher overall levels. Second, the figure also shows some heteroscedasticity, that is uneven variance, across the countries. Note, for example, Romania with little variance, compared with Lithuania, with much greater variance.

Next, let's run an analysis of variance to assess more formally whether this data suffers from unit effects:

```
anova<-aov(unempl~cid, data=D) #run ANOVA of unemployment by unit variable 'cid'
summary(anova) #view results
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## cid          10   1564   156.4      17 <2e-16 ***
## Residuals   199   1831     9.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA results above show a very low p-value, suggesting significant unit effects. This should not be surprising, given our data. What this means is that there is important variation in unemployment across our eleven countries. This suggests that regular OLS analysis will be biased!

For the sake of the excercies, let's estimate an OLS model, predicting unemployment `unempl` as a function of GDP per capita `gdp.cap`, government education expenditure `exp.ed`, and external remittances `remit`:

```
ols<-lm(unempl~gdp.cap+exp.ed+remit, data=D)
summary(ols)
```

```
##
## Call:
## lm(formula = unempl ~ gdp.cap + exp.ed + remit, data = D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0556 -2.4639 -0.1195  1.7877  9.4412
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.250e+01  1.539e+00   8.121 4.20e-14 ***
## gdp.cap     -3.351e-04  4.172e-05  -8.032 7.29e-14 ***
## exp.ed       7.155e-01  2.872e-01   2.491   0.0135 *
## remit        6.583e-01  1.503e-01   4.379 1.90e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.442 on 206 degrees of freedom
## Multiple R-squared:  0.281,  Adjusted R-squared:  0.2705
## F-statistic: 26.83 on 3 and 206 DF,  p-value: 1.082e-14
```

This looks like a nice model… but, with the prior knowledge of unit effects, we should not be too excited. This model is biased. How can we deal with the presence of unit effects, and estimate the influence of our three predictors without bias?

There are two general approaches to dealing with unit effects:

1) Fixed effects models
2) Random effects models

Fixed Effects Models

There are two ways to estimate fixed effects, and both lead to the same outcome. Let us start with a generic TSCS equation:

$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + a_i + u_{it}$

Here x_{it} represents predictors that vary over time and across units, as the subscripts suggest.

z_i represents predictors that vary across units only. These are called time-invariant variables.

The overall residual, normally denoted ϵ_{it}, is split into two parts in TSCS. In the equation above, a_i is the unobserved variance of the dependent variable y_{it} that varies across units, but not over time. It captures unit effects. u_{it} is idiosyncratic error that varies over time and across units.

Fixed Effects Estimator

One way to arrive at fixed effects estimation and deal with unit effects is by subtracting the mean values of the variables across countries over time. Consider the TSCS regression equation again:

$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + a_i + u_{it}$ (1)

Taking the mean of this equation over time produces this:

$\bar{y}_{it} = \beta_1 \bar{x}_{it} + \beta_2 z_i + a_i + \bar{u}_{it}$ (2)

(Note that the over-time mean of z_i is z_i, and of a_i is a_i because they are both time-invariant.)

Now, if we subtract equation (2) from equation (1), we get the so-called *time-demeaned equation*:

$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i$ (3)

This equation (3) is the *fixed effects estimator*, also known as *within estimator* because it estimates over-time effects *within* units.

Dummy Variable Estimator

The second way of estimating fixed effects is by doing the following:

$y_{it} = \beta_0 + \beta_1 x_{it} + \gamma C_i + u_{it}$ (4)

Here C_i are N dummy variables = 1 for unit i, otherwise 0. We are effectively modelling the unit effects as 'country dummies.' This is known as the *dummy variable estimator*. Note that this produces the same results as the fixed effects estimator.

Estimating Fixed Effects in R

Let's go back to our eastern Europe data, and reestimate it using fixed effects:

```
fe<-plm(unempl~gdp.cap+exp.ed+remit, data=D, index=c("cid","year"), model="within")
  #define model, data, unit and time variables, 'within' suggests *fixed effects estimator*
summary(fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = unempl ~ gdp.cap + exp.ed + remit, data = D, model = "within",
##     index = c("cid", "year"))
##
```

```
## Unbalanced Panel: n = 11, T = 9-23, N = 210
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -7.0658855 -1.7947750 -0.0056007  1.5909313  7.1275419
##
## Coefficients:
##            Estimate  Std. Error t-value  Pr(>|t|)
## gdp.cap -3.5468e-04  5.0641e-05 -7.0039 3.886e-11 ***
## exp.ed   8.0952e-02  4.0927e-01  0.1978  0.843409
## remit    5.8909e-01  1.9372e-01  3.0410  0.002681 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1830.8
## Residual Sum of Squares: 1446.6
## R-Squared:      0.20984
## Adj. R-Squared: 0.15743
## F-statistic: 17.3506 on 3 and 196 DF, p-value: 4.9503e-10
```

This model has removed unit effects. Note that GDP becomes slightly stronger, while education expenditure loses significance in comparison with the OLS model we estimated earlier. Let's compare this estimate with a *dummy variable* model, which is estimated simply by running an OLS model and including country dummies:

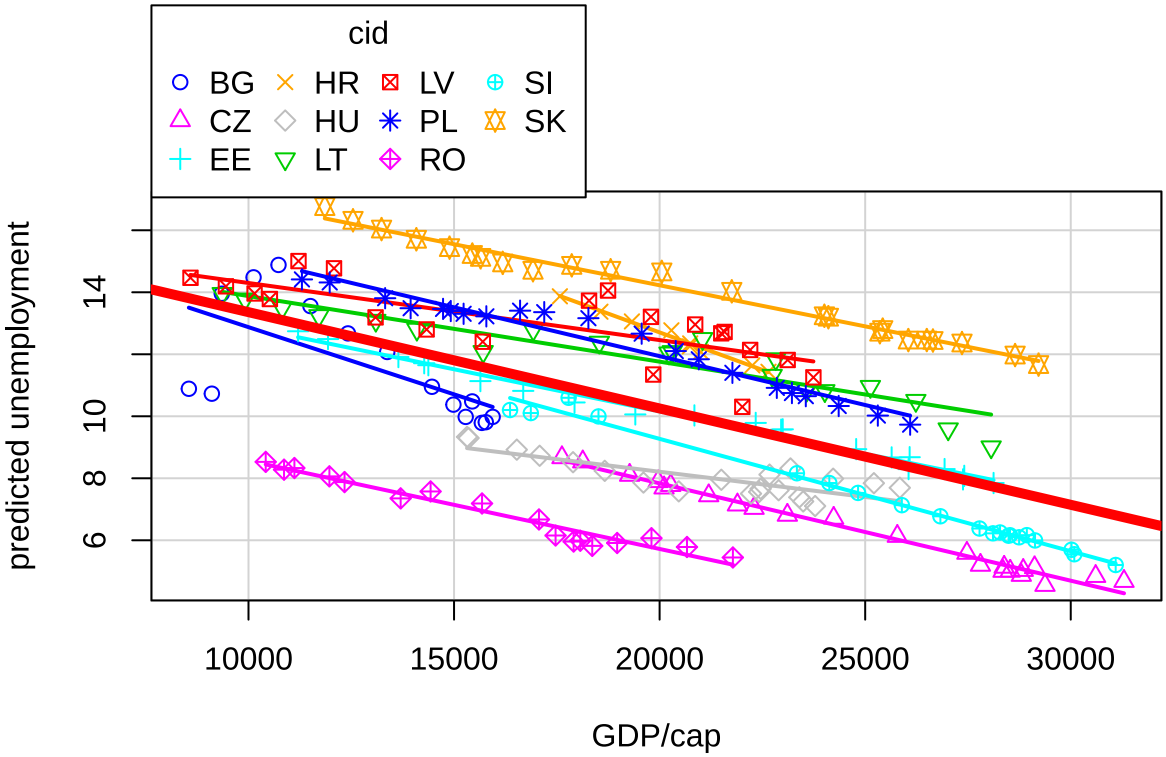
```
fe.dummy<-lm(unempl~gdp.cap+exp.ed+remit+factor(cid)-1, data=D)
summary(fe.dummy)
```

```
##
## Call:
## lm(formula = unempl ~ gdp.cap + exp.ed + remit + factor(cid) -
##     1, data = D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0659 -1.7948 -0.0056  1.5909  7.1275
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## gdp.cap       -3.547e-04  5.064e-05  -7.004 3.89e-11 ***
## exp.ed         8.095e-02  4.093e-01   0.198  0.84341
## remit          5.891e-01  1.937e-01   3.041  0.00268 **
## factor(cid)BG  1.354e+01  1.817e+00   7.450 2.91e-12 ***
## factor(cid)CZ  1.438e+01  2.297e+00   6.260 2.37e-09 ***
## factor(cid)EE  1.621e+01  2.623e+00   6.179 3.67e-09 ***
## factor(cid)HR  1.740e+01  2.199e+00   7.913 1.81e-13 ***
## factor(cid)HU  1.420e+01  2.390e+00   5.940 1.28e-08 ***
## factor(cid)LT  1.688e+01  2.396e+00   7.047 3.03e-11 ***
## factor(cid)LV  1.669e+01  2.493e+00   6.693 2.24e-10 ***
## factor(cid)PL  1.777e+01  2.415e+00   7.359 4.97e-12 ***
## factor(cid)RO  1.184e+01  1.809e+00   6.543 5.13e-10 ***
## factor(cid)SI  1.541e+01  2.737e+00   5.633 6.10e-08 ***
## factor(cid)SK  2.033e+01  2.090e+00   9.727  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.717 on 196 degrees of freedom
## Multiple R-squared:  0.9433, Adjusted R-squared:  0.9392
## F-statistic: 232.7 on 14 and 196 DF,  p-value: < 2.2e-16
```

Note that the estimates from the dummy variable model are identical to those of the fixed effects estimator above. The difference is that the dummy variable model also estimates the γ coefficients from equation (4) – these are effectively the individual country intercepts. We can approach their visualization in the following way:

```
yhat <- fe.dummy$fitted #take predicted values
scatterplot(yhat~gdp.cap | cid , data=D, boxplots=FALSE, xlab="GDP/cap", ylab="predicted unemployment",
  #scatter predicted values by GDP and cid
abline(lm(unempl~gdp.cap, data=D),lwd=5, col="red")
```



The above figure demonstrates the logic of unit effects – we can see that different counties have different levels of unemployment. The effects are the same (note, they appear to differ a bit because we are visualizing a multi-dimensional regression plane in two dimensions), but they have different intercepts, which we estimated with the dummy variable model. The thick red line is the simple (OLS) relationship.

Problems with fixed effects estimation

Fixed effects (FE) is a powerful method of dealing with unit effects, but it has a number of flaws. First, it is inefficient because it estimates many parameters – either because it estimates country intercepts or because it subtracts means.

Second, fixed effects is a brutal way of dealing with unit effects by simply washing them away. It takes the error term ϵ_{it}, which can be considered as "general ignorance," and assigns a part of it to units (a_i). But this is still ignorance because we have no theoretical explanation for it. It is what Maddala called "specific ignorance." It is typical of social science approaches that do not care for unit (country) distinctions, and just model it away.

Third, fixed effects are catastrophic for social scientists who care about time-invariant variables, such as institutions… As equation (3) above suggests, these variables are thrown out with fixed effects because they do not vary over time.

Finally, fixed effects cannot estimate level effects, explaining the actual levels reached by the dependent variable, as levels are removed in the subtraction of the mean, or the estimation of country dummies.

Random Effects Models

Random effects (RE) models help solve some of the problems of fixed effects. Random effects, like in multi-level models, generally assume that time observations are nested within units. The random effects approach views unit effects a_i as ignorance which is illogical to model, and belongs to the residuals. Random effects thus proceed from:

$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \epsilon_{it}$ (5)

Where the residual $\epsilon_{it} = a_i + u_{it}$, thus including unit effects.

However, moving unit effects into the residual introduces two problems. First, is serial autocorrelation because a_i is in the error at each time point t. The second is that if a_i is correlated with the predictors x_{it}, then we have omitted variable bias. We can handle these by transforming the error to account for serial autocorrelation. Specifically, use the knowledge of the size of error variance caused by unit effects a_i. Concerning the second problem, well, we can wave our hands, and … assume that $Cov(a_i, x_{it}) = 0$. This, however, is a very, very, very strong assumption.

Random Effects Estimator

The random effects estimator is then derived from the following transformation. Going back the equation (5):

$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \epsilon_{it}$ (5)

We next define a transformation coefficient λ:

$\lambda = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2}}$

The transformation of equation (5) is then:

$y_{it} - \lambda\bar{y}_i = \beta_0(1 - \lambda) + \beta_1(x_{it} - \lambda\bar{x}_i) + \beta_2(z_i - \lambda z_i) + (\epsilon_{it} - \lambda\bar{\epsilon}_i)$ (6)

This transformed *quasi-demeaned* equation (6) is similar to the *demeaned* equation (3), only the means are multiplied by λ. The random effects transformation subtracts a fraction of the time average, where the fraction depends on σ_u, σ_a and T. This means that we are using our knowledge of the *within* error variance σ_u^2 (estimated with fixed effects) and the total variance σ^2 (estimated with OLS). Note that time-invariant predictors z_i are now possible.

Given the λ transformation, the following is the case: When $\lambda = 1$, the RE model is identical to an FE model. When $\lambda = 0$ the RE model is identical to OLS estimation. That means that the bigger the variance of the unobserved effect, the closer RE is to FE, the smaller the variance of the unobserved effect, the closer RE is to OLS.

Estimating Random Effects in R

Let's reestimate our previous model using random effects.

```
re<-plm(unempl~gdp.cap+exp.ed+remit, data=D, index=c("cid","year"), model="random")
  #same logic as FE estimation above, but "random" specifies RE estimator
summary(re)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = unempl ~ gdp.cap + exp.ed + remit, data = D, model = "random",
##     index = c("cid", "year"))
##
## Unbalanced Panel: n = 11, T = 9-23, N = 210
##
## Effects:
##                  var std.dev share
## idiosyncratic 7.381   2.717 0.526
## individual    6.651   2.579 0.474
## theta:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6687  0.7707  0.7760  0.7673  0.7760  0.7855
##
## Residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -6.8140 -1.8684 -0.2758  0.0069  1.7225  7.1544
##
## Coefficients:
##                 Estimate  Std. Error z-value  Pr(>|z|)
## (Intercept) 15.18007915  2.16610393  7.0080 2.417e-12 ***
## gdp.cap     -0.00034885  0.00004746 -7.3504 1.976e-13 ***
## exp.ed       0.20766429  0.38119247  0.5448  0.585908
## remit        0.58907787  0.18099659  3.2546  0.001135 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1938.9
## Residual Sum of Squares: 1502.8
## R-Squared:      0.22492
## Adj. R-Squared: 0.21363
## Chisq: 59.7719 on 3 DF, p-value: 6.5763e-13
```

These results are quite similar to our fixed effects model above, albeit with some minor differences. It is important to note that random effects coefficients are a combined within unit and over time effects, and between unit effects. The βs in random effects estimates thus represent the average effect of x over y when x changes across time and between countries by one unit.

Fixed or random effects?

One way to check whether FE is preferred to RE or vice versa is to run the Hausman test. Here H_0 is that the preferred model is RE. The alternative H_a is that it is FE. We can run this test on our saved models:

```
phtest(fe,re)
```

```
##
##  Hausman Test
##
## data:  unempl ~ gdp.cap + exp.ed + remit
## chisq = 0.91686, df = 3, p-value = 0.8214
## alternative hypothesis: one model is inconsistent
```

This effectively tests the (very very very strong) assumption that the errors a_i are uncorrelated with the predictors ($Cov(a_i, x_{it}) = 0$). If the $p - value > 0.05$, the assumption does not hold. We should then use fixed effects. This is our case here, so our fixed effects model is preferable.

Panel Corrected Standard Errors

As we have seen, both FE and RE have various problems. The big concern with FE is its inability to estimate effects of time-invariant variables. The big concern with RE is its strong assumption that unit effects are uncorrelated with the predictors. Seeking an alternative estimation technique, Beck and Katz (1995) proposed estimating TSCS data using regular OLS, but correcting the standard errors in order to make them robust to contemporaneous error correlation and autocorrelation. They called this error correction *Panel Corrected Standard Errors* (PCSE).

Note that, as the name suggests, PSCE corrects the standard errors, which has no impact on the estimated coefficients, only on their significance tests. Consequently, PCSE coefficients are the same as OLS coefficients, only the standard errors differ.

Estimating PCSE models

In R, we can estimate PCSE using the `pcse` library. We first estimate the model using OLS, then, we recalculate the standard errors to get the PCSE error estimates:

```r
library(pcse)
ols<-lm(unempl~gdp.cap+exp.ed+remit, data=D) #estimate OLS model
pcse<-pcse(ols, groupN=D$cid, groupT=D$year)
  #calculate PCSE, identifying the original OLS model, and unit and time variables
summary(pcse)
```

```
##
##  Results:
##
##                  Estimate         PCSE   t value     Pr(>|t|)
## (Intercept) 12.4951714866 1.579019e+00  7.913247 1.517870e-13
## gdp.cap     -0.0003350939 5.819013e-05 -5.758604 3.051391e-08
## exp.ed       0.7155133869 2.727566e-01  2.623267 9.360468e-03
## remit        0.6582911363 1.811674e-01  3.633607 3.529055e-04
##
##  ---------------------------------------------
##
## # Valid Obs = 210; # Missing Obs = 65; Degrees of Freedom = 206.
```

Note that these results are very similar to the OLS model we estimated at the very beginning. The coefficients are the same, the errors differ, but the hypothesis test conclusions are the same – all three predictors have significant effects. This is in contrast to our FE and RE models, in which educational expenditure had an insignificant effect.

Problems in PCSE models and lagged dependent variable as a solution

The important thing to remember is that by correcting standard errors, PCSE models can deal only with issues in the errors, and specifically contemporaneous error correlation. This is a relatively minor problem, and can be dealth with by modeling it (for example, we can include a dummy variable for the year 2020, to capture the contemporaneous shock of the COVID crisis). PCSE cannot deal with the big problem, potentially biasing the estimates – unit effects.

In order to deal with the big problem of unit effects, Beck and Katz propose including a *lagged dependent variable* (LDV) in the OLS model before calculating the PCSE:

$y_{it} = \beta_0 + \alpha y_{t-1} + \beta_1 x_{it} + \beta_2 z_i + \epsilon_{it}$ (7)

Here y_{t-1} is the lagged dependent variable, and appears on the right hand side of the equation.

The inclusion of LDV is a reasonable proposal. The LDV models history of the dependent variable y, and thus deals with two things at once. First, it models the temporal dynamics of y, particularly autoregression. Second, since it models history of y in each unit, it captures unit effects.

It is, however, important to note that the control for unit effects via the LDV will be as strong as the dynamics in the series. That is to say, if there are weak dynamics in the series y, and the coefficiant on the LDV (α in equation 7) is small, say < 0.5, then the control for unit effects will be similarly small. Therefore, we can feel comfortable about handling unit effects with the LDV only as the LDV coefficient α gets closer to 1.

Let's first create a LDV within each of our country panels:

```r
library(dplyr)
D %>% #from D dataframe
  group_by(cid) %>% #by cid
  mutate(l.unempl=dplyr::lag(unempl)) %>% #create lag of unemployment
  mutate(d.unempl=unempl-dplyr::lag(unempl)) %>% #create differences of the other variables
  mutate(d.gdp.cap=gdp.cap-dplyr::lag(gdp.cap)) %>%
  mutate(d.exp.ed=exp.ed-dplyr::lag(exp.ed)) %>%
  mutate(d.remit=remit-dplyr::lag(remit)) -> #assign the result to D
  D
head(D) # view data
```

```
## # A tibble: 6 x 12
## # Groups:   cid [1]
##   country cid   year  gdp.cap unempl exp.ed remit l.unempl d.unempl d.gdp.cap
##   <fct>   <fct> <fct>   <dbl>  <dbl>  <dbl> <dbl>    <dbl>    <dbl>     <dbl>
## 1 Bulgar~ BG    1996    8550.   13.5   2.25 0.341     NA      NA          NA
## 2 Bulgar~ BG    1998    9109.   12.2   2.76 0.339     13.5    -1.31      559.
## 3 Bulgar~ BG    2001    9351.   19.9   3.41 5.87      12.2     7.72      242.
## 4 Bulgar~ BG    2002   10123.   18.1   3.41 7.23      19.9    -1.81      773.
## 5 Bulgar~ BG    2003   10730    13.7   4.03 8.19      18.1    -4.38      607.
## 6 Bulgar~ BG    2004   11507.   12.0   2.34 6.64      13.7    -1.70      777.
## # ... with 2 more variables: d.exp.ed <dbl>, d.remit <dbl>
```

```r
C<-na.omit(D) #remove missing rows (first in each unit)
```

Let's include a LDV in our PCSE analysis:

```r
ols.ldv<-lm(unempl~l.unempl+gdp.cap+exp.ed+remit, data=C) #estimate OLS model with LDV
pcse.ldv<-pcse(ols.ldv, groupN=C$cid, groupT=C$year)
  #calculate PCSE
summary(pcse.ldv)
```

```
##
##  Results:
##
##                  Estimate         PCSE    t value     Pr(>|t|)
## (Intercept)  2.410976e+00 1.489794e+00  1.6183288 1.072164e-01
## l.unempl     7.936007e-01 7.096955e-02 11.1822702 1.002012e-22
## gdp.cap     -8.419179e-05 5.380237e-05 -1.5648344 1.192515e-01
## exp.ed       2.373898e-01 2.392310e-01  0.9923040 3.222853e-01
## remit        8.261295e-02 1.638437e-01  0.5042181 6.146803e-01
```

```
##
## --------------------------------------------
##
## # Valid Obs = 199; # Missing Obs = 65; Degrees of Freedom = 194.
```

The inclusion of the LDV clearly updated the estimates of our three predictors of interest. The LDV coefficient, $\alpha = 0.79$, suggests that there are significant temporal dynamics in our data. Thinking deeper about this, we should realize a major error we have been committing all along! It is possible that our time-series are integrated. Non-stationary time series may cause a number of problems, particularly if the non-stationarity comes from trends in the data!

Estimating Integrated TSCS

In fact, before estimating any TSCS, we should consider the dynamics of our time-series across all the panels. The key concerns should be trending. If our dependent variable is trended, and any predictor is trended within any panel, these will naturally correlate, whether or not they are actually causally related. This will bias our analysis.

Dickey Fuller test in TSCS

For this reason, before running any TSCS, we should check our data for stationarity using the Dickey Fuller test. However, this is more tricky than in simple time-series, we want to see the dynamics of our dependent variable, as well as our key predictors, within each panel. To do this most effectively, let's run a loop that will calculate the Dickey Fuller test by each country, and construct a data frame with the test results, which I will call `DF`:

```
D$c<-as.numeric(as.factor(D$cid)) #create a counter for each country 'c'
DF<-matrix(data=NA, nrow=11, ncol=5) #create an empty matrix to be filled with test p-values

#for loop testing the key variables in each country 'c'
for (i in 1:11) {
DF[i,1]<-i
DF[i,2]<-adf.test(D$unempl[D$c==i])$p.value
DF[i,4]<-adf.test(D$gdp.cap[D$c==i])$p.value
DF[i,3]<-adf.test(D$exp.ed[D$c==i])$p.value
DF[i,5]<-adf.test(D$remit[D$c==i])$p.value
}
```

```
## Warning in adf.test(D$gdp.cap[D$c == i]): p-value smaller than printed p-value
```

```
## Warning in adf.test(D$remit[D$c == i]): p-value greater than printed p-value
```

```
## Warning in adf.test(D$unempl[D$c == i]): p-value smaller than printed p-value
```

```
#name variables
colnames(DF) <- c("cid","unempl","gdp.cap","exp.ed","remit")

#view frame
DF
```

```
##       cid    unempl    gdp.cap    exp.ed     remit
##  [1,]   1 0.9179881 0.63377380 0.9557575 0.1031596
##  [2,]   2 0.6080650 0.84652517 0.7160169 0.3888277
##  [3,]   3 0.4626505 0.43470209 0.8218280 0.7206044
##  [4,]   4 0.9561627 0.06608503 0.0100000 0.7537133
##  [5,]   5 0.5230235 0.48053065 0.5150876 0.3615191
##  [6,]   6 0.5177776 0.07253250 0.6514034 0.9900000
##  [7,]   7 0.3019558 0.59052933 0.7409686 0.6465403
##  [8,]   8 0.0100000 0.21931081 0.3558345 0.5986065
##  [9,]   9 0.1713443 0.84999738 0.4469987 0.3448619
## [10,]  10 0.4691414 0.74288435 0.6676331 0.6452202
## [11,]  11 0.5517866 0.35305047 0.6311971 0.3294069
```

As we can see in the `DF` frame above, the p-values for the Dickey Fuller tests across our key variables and the countries are greater than 0.05 in almost all of the cases. This means that we have integration (almost) across the board. In a situation like this, the safest way to proceed is to consider a difference model.

Difference models as a solution to integrated series

A difference model simply models not the levels (values) of the variables at hand, but rather their differences. Instead of testing whether the level of unemployment is a function of the level of GDP per capita etc., we rather test whether the change in unemployment is a function of the change in GDP per capita. Looking at changes removes the integration and trend in time-series.

It is important to note that this holds as long is the integration is of the first order I(1). In case the integration is of a second order (say, an exponential trend), then we should take the second differences (difference of the difference) and so on. Mathematically, a difference model looks like this:

$\Delta y = \beta_0 + \beta_1 \Delta x1_{it} + \beta_2 \Delta x2_{it} + ... + \epsilon_{it}$

Where Δy is the first difference in y: $\Delta y = y_t - y_{t-1}$ and so on.

Let's reestimate our analysis using a difference model. We can apply this simply by running OLS regression with differenced variables – note that we created these earlier in the lesson, and denoted them with `d.`. Alternatively, we can run the 'first difference' model in the `plm` package, by specifying `model="fd"`:

```
#OLS model with differenced variables
ols.d<-lm(d.unempl~d.gdp.cap+d.exp.ed+d.remit, data=C)
summary(ols.d)
```

```
##
## Call:
## lm(formula = d.unempl ~ d.gdp.cap + d.exp.ed + d.remit, data = C)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3524 -0.8597 -0.2429  0.8061  5.1110
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6785073  0.1430052   4.745 4.03e-06 ***
## d.gdp.cap   -0.0013066  0.0001174 -11.130  < 2e-16 ***
## d.exp.ed    -0.0428618  0.2543895  -0.168 0.866373
## d.remit      0.5961188  0.1734021   3.438 0.000717 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.609 on 195 degrees of freedom
## Multiple R-squared:  0.4181, Adjusted R-squared:  0.4092
## F-statistic: 46.71 on 3 and 195 DF,  p-value: < 2.2e-16
```

```
#First difference model (plm package)
fe.d<-plm(unempl~gdp.cap+exp.ed+remit, data=D, index=c("cid","year"), model="fd") #"fd" specifies first
summary(fe.d)
```

```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = unempl ~ gdp.cap + exp.ed + remit, data = D, model = "fd",
##     index = c("cid", "year"))
##
## Unbalanced Panel: n = 11, T = 9-23, N = 210
## Observations used in estimation: 199
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -5.35241 -0.85966 -0.24289  0.80615  5.11098
##
## Coefficients:
##                 Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)  0.67850732  0.14300518   4.7446 4.031e-06 ***
## gdp.cap     -0.00130656  0.00011739 -11.1304 < 2.2e-16 ***
## exp.ed      -0.04286177  0.25438948  -0.1685 0.8663734
## remit        0.59611878  0.17340205   3.4378 0.0007169 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    867.84
## Residual Sum of Squares: 504.96
## R-Squared:      0.41814
## Adj. R-Squared: 0.40919
## F-statistic: 46.7109 on 3 and 195 DF, p-value: < 2.22e-16
```

The above models produce essentially identical results (with minor differences only in the small decimals). Note that the differencing can remove unit effects that were present in the variable levels. This is because we are no longer looking at the levels of variables (that differ due to historical idiosyncracies), but at the change from one period to the next. However, if units have different rates of change, such as differential GDP growth rates, then unit effects may remain. This would effectively mean that the variable is integrated at a higher order than 1. The way to check whether this is the case, is to run a diThe models above suggest GDP and remissions are significant predictors of unemployment, but education expenditure is not.

Heteroscedasticity

As mentioned above, TSCS may suffer from heteroscedasticity. While this can be the case in any dataset, in TCSE, heteroscedasticity can be caused by the naturally divergent variances across the units. To test for heteroscedasticity we can apply the Breusch-Pagan test, here focusing on our difference model:

```
bptest(d.unempl~d.gdp.cap+d.exp.ed+d.remit, data=C)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  d.unempl ~ d.gdp.cap + d.exp.ed + d.remit
## BP = 11.659, df = 3, p-value = 0.008649
```

Note that the H_0 for the Breusch-Pagan test is homoscedasticity, thus we want to fail to reject it (we want a higher p-value). Here, we clearly reject the null and conclude the presence of heteroscedasticity.

Heteroscedasticity biases our standard error estimates, and thus potentially leads to incorrect significance tests. A remedy for heteroscedasticity is an alternative calculation of our standard errors. We can re-estimate our standard errors by applying the heteroscedasticity-consistent variance estimators.

```
coeftest(ols.d, vcovHC) #recalculate SE from OLS difference model
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.6785073  0.1836329  3.6949 0.0002857 ***
## d.gdp.cap   -0.0013066  0.0001976 -6.6123 3.544e-10 ***
## d.exp.ed    -0.0428618  0.2576805 -0.1663 0.8680642
## d.remit      0.5961188  0.4172843  1.4286 0.1547284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the corrected standard error calculation, we can see that only GDP per capita is a significant predictor of unemployment.